Translating English Verbal Collocations into Spanish: on Distribution and other Relevant Differences related to Diatopic Variation

Gloria Corpas-Pastor
University of Malaga (Spain) / University of Wolverhampton (United Kingdom)

Introduction

Collocation as a pervasive phenomenon in language refers to the tendency of words to occur together and exhibit idiosyncratic combinatory properties. Collocation is also the term which denotes the resulting word combinations. For instance, within the legal domain typical sequences are *reckless driving*, *to enact a law* or *breach of contract* (*conducción temeraria*, *promulgar una ley*, *incumplimiento de contrato* in Spanish). They are salient word combinations that not only denote the legal field, but also convey conventionalised ways of expressing a state of affairs or an action. Collocations tend to be semantically transparent, but not always predictable. Unlike idioms (eg. *to carry coals to Newcastle* or *a la chita callando*, 'on the sly') collocations do not normally pose problems in comprehension, but mainly in production. Collocational non-isomorphism is particularly noticeable in contrastive studies and translation. This applies to general language and domain-specific registers alike. For example, in English *hardened* collocates with *bachelor*, but not with *drunker* or *smoker;* whereas the equivalent adjective *empedernido* does not have the same restrictions (*fumador/soltero/borracho empedernido*). Similarly, in the legal/administrative domain it is typical to say *to file/lodge an appeal*, but not *to send an appeal; and similarly in Spanish, the verbal collocates for *recurso* ('appeal') are *presentar* and *interponer*, whereas *enviar (send) is not a plausible verbal collocate.

Combinatory restrictions reflect a language's idiosyncrasy in general but also as regards registers, levels of formality and language varieties (diastratic, diaphasic and diatopic restrictions). For example, the English verbal collocation *take away parental rights* classes as general language, neutral, whereas *terminate parental rights* is marked as specialised legal English, formal; in Spanish 'to give someone a fine' would require different verbal collocates depending on the degree of formality and specialisation: *poner una multa* (general language, neutral), *imponer una multa* (specialised legal/administrative Spanish, formal), and *cascar una multa* (general language, informal) (Corpas Pastor, 2015). Language varieties also reflect their own peculiarities as regards collocability. Quirk et al. (1989, p. 752) claim that collocations with alternative delexical verbs (*have/take a break*) tend to select *take* in American English and *have* in British English. Similar differences of usage can be observed in Spanish: 'catch a cold' is *coger un resfriado* (Spain) or *pescar/pillar un resfriado* (Chile and Mexico) (Molero,

2003).

In translation, the rendering of collocations into the target language is usually governed by the languages' anisomorphism at the level of lexical selection of collocates. Though bases are usually translated literally, collocates do not seem to follow this straightforward path: *to pay homage* cannot be translated into Spanish as *pagar homenaje, but prototypically as *rendir homenaje*.[1] However, in the case of widespread transnational languages, like Spanish and English, things get more complicated. The crucial question then becomes not so much which translation equivalent in the "target language" to choose as the actual "target language variety" which is at stake.

Bilingual dictionaries tend to favour a simplified, prototypical approach to collocations. For example, Collins[2], Larousse[3] and Oxford[4] dictionaries only provide *rendir homenaje* as possible translation equivalent (see the entry for *homage* in CSD, LSD and OSD). A similar situation applies as regards actual translations. The most common translation equivalent is *rendir homenaje*, although there appear some transpositions (e.g. *en homenaje*) and modulations (e.g. *tributo* instead of *homenaje*), as seen in the examples below (1-3), extracted from Linguee[5].

(1) [EN] *Speaking before the House today, I should like to* ***pay homage*** *to the Polish bishops*.

[ES] *Al hablar hoy ante la Cámara, quisiera* ***rendir homenaje*** *a los obispos polacos*.

(2) [EN] *Ladies and gentlemen, I would like to ask you to observe a minute's silence to* ***pay homage*** *to all the victims*.

[ES] *Les rogaría, si están de acuerdo Señorías, guardar un minuto de silencio* ***en homenaje*** *a todas estas víctimas*.

(3) [EN] *In concluding these remarks at […] this Conference held to mark the fiftieth anniversary of UNESCO, it is fitting to* ***pay homage*** *to the far-sighted vision of the Organization's founders*.

[ES] *Como conclusión de estos comentarios en esta Conferencia organizada para celebrar el cincuentenario de la UNESCO, corresponde* ***rendir tributo*** *a la amplia visión de los fundadores de la Organización*.

---

[1] The verb *rendir* is the most frequent and salient collocate for *homenaje* in the esTenTen [2011] corpus of general Spanish (cf. 3.1.).

[2] *Collins Spanish-English Dictionary* (On-line version). [CSD]
<http://www.collinsdictionary.com/dictionary/english-spanish>
<http://www.collinsdictionary.com/dictionary/spanish-english

[3] *Larousse Spanish-English Dictionary* (On-line version). [LSD]
< http://www.larousse.com/es/diccionarios/ingles-espanol/>
<http://www.larousse.com/es/diccionarios/espanol-ingles>

[4] *Oxford Spanish-English Dictionary* (On-line version). [OSD]
<http://www.oxforddictionaries.com/spanish/>

[5] http://www.linguee.es/. [Accessed: 12 June 2015].

However, general Spanish[6] offers a wider range of collocates in a cline of formality and grammaticalisation, the most frequent and salient ones being *dar/hacer un homenaje* (informal/neutral; delexical collocational senses) and *rendir/tributar/brindar un homenaje* (formal; figurative, coerced collocational senses). And yet, the former list seems to be further restricted by diatopic considerations. A few examples will suffice: *dar* and *brindar* are not plausible verbal collocates for *homenaje* in Nicaraguan Spanish; Mexican and Argentinian Spanish do not use *dar* but *hacer* as delexical verb with *homenaje*; Dominican Spanish only permits collocations with *rendir* and *tributar*, but not *brindar*, *dar* nor *hacer;* in Honduran Spanish the only option is *rendir*, etc.[7]

We argue that language varieties should be taken into account in order to enhance fluency and naturalness of translated texts. In this paper we will examine the collocational verbal range for prima-facie translation equivalents of words like *decision* and *dilemma*, which in both languages denote the act or process of reaching a resolution after consideration, resolving a question or deciding something. Restricting the choice of nouns this way allows us to place special emphasis on semantic-functional counterparts and their collocational preferences. We will be mainly concerned with diatopic variation in Spanish. To this end, we set out to develop a giga-token corpus-based protocol which includes a detailed methodology sufficient to detect collocational peculiarities of transnational languages and which is easily reproducible by researchers. To our knowledge, this is one of the first observational studies of this kind.

The paper is organised as follows. Section 2 provides a feature characterisation of collocations. Section 3 deals with the choice of corpora, corpus tools, nodes and patterns. Section 4 covers the automatic retrieval of the selected verb + noun (object) collocations in general Spanish and the co-existing national varieties. Special attention is paid to comparative results in terms of similarities and mismatches. Section 5 presents conclusions and outlines avenues of further research.

2. Defining collocations

The term *collocation* was introduced by Firth (1957, 1968) not only to mean a mode of semantic analysis (meaning by collocation), but also a stylistic means to characterise restricted languages and levels of formality. Firth's notion of collocation "the habitual company a key-word keeps" (Firth, 1968: 113) makes reference to usual or habitual co-occurrence, i.e. restricted (or preferred) lexical selection. More than half a century later, there is still very little consensus on the nature of collocations nor on their distinctive features. A suitable working definition would probably revolve around concepts such as lexical selection,

[6] Data extracted from the esTenTen corpus (cf. 3.1).
[7] Data extracted from the esAmTenTen subcorpora (cf. 3.1).

semantic cohesion, syntactic relationship, frequency, recurrence, salience and institutionalisation.

A full discussion on the notion of collocation and competing theories is clearly beyond the scope of this paper (cf. Corpas Pastor, 1996, 2001, 2015; Bartsch, 2004; Barnbrook, Krishnamurthy and Mason, 2013). In what follows we will merely attempt to characterise collocation in terms of features as a convenient background to our analyses.

1.1. Lexical restriction and variation

Syntagmatic and paradigmatic lexical restriction is core to the definition of collocations. According to Cowie (1981), collocations are composite units that allow substitution of at least one of its components without semantic change of the other ones: e.g. *to freeze wages* (*congelar el sueldo),* where *wages* could be replaced with *prices* or *income* (and *sueldo* with *salario* or *precio*) while *to freeze* and *congelar* keep their actual figurative or specialised sense ('fix at a given or current level'); or *to explode a myth (desmontar un mito),* where *to explode* and *desmontar* in the sense of 'show something to be false or no longer true', can only combine with some words denoting 'misconception': e.g. *myth, belief, idea, notion* or *theory*, in English; and *falacia, estereotipo, prejuicio* or *mentira,* in Spanish.

While restricted lexical selection means combinatory idiosyncrasy, it is a well-known fact that collocations are not only peculiar to a given language, but are also one of the more powerful means to characterise style. Corpus stylistics exploits the distributional properties of words (including collocations) to identify features that are characteristic of a particular text (literary style) or an author (authorial style).[8] An example of the former is Hardy (2004, 2007), who studies literary style by means of interlingual collocational analysis; an example of the latter is Hoover (2003), who found that cluster analysis based on frequent collocations provides a robust and more accurate method of authorship attribution than analyses just based on either words or sequences. Corpus-based register analyses have also established that domain-specific genres and restricted registers tend to exhibit distinctive collocational patterns (Biber and Conrad, 1999; Gledhill, 2000; Williams, 2002, etc.). Corpas Pastor (2015) provides examples of collocations for 'giving someone a fine' which evidence a cline of register-specificity and formality restriction, such as *take away parental rights* (general language, neutral) versus *terminate parental rights* (specialised legal English, formal); and *poner una multa* (general language, neutral) versus *imponer una multa* (specialised legal/administrative Spanish, formal) and *cascar una multa* (general language, informal).

Collocations also help to characterise language change and linguistic variation, as evidenced by Hilper (2006), one of the first authors to study

[8] For a brief overview on corpus-based stylistics and collocational analysis, see Biber (2011).

collocational lexical change over time; and Torres Cacoullos and Walker (2011), who compare collocations in grammaticalisation and linguistic variation for Spanish and English, among many others. The same applies to collocational restrictions according to language varieties. For example, Greenbaum (1974) reported usage differences among verb-intensifier collocations: e.g., in British English, *entirely* collocates with verbs denoting 'agreement/disagreement', while *completely* tends to collocate with verbs of 'failure', whereas in American English this distinction is blurred. In the case of Spanish, there are collocational diatopic differences among the national varieties: e.g., 'switch on the light' is *encender la luz* (Spain) or *prender la luz* (Chile), and 'brush one's teeth' is *limpiarse los dientes* (Northern Spain) or *lavarse los dientes* (Southern Spain, the Canary Islands and Latin America) (Koike, 2001).

1.2. Semantic and syntactic boundedness

Contrary to idioms, collocations are not fixed and their meanings are essentially compositional (cf. section 1). However, collocations exhibit a certain degree of internal cohesion and institutionalisation. Collocations help to disambiguate polysemous items, convey tipicality or select particular senses, i.e. delexical, specialised, figurative, coerced (Corpas Pastor, 2015). As a matter of fact, meaning has proved crucial in the shaping of the notion of collocation. Meaning by collocation is essentially a corpus-based distributional model of linguistic analysis which strives to statistically uncover significant word co-occurrences. A crucial feature is the sense relations that exist between the constituents of a given collocation (semantic cohesion or boundedness). Collocations can be conceived as a bipartite structure, recursive and conventionally restricted, in which both collocates exhibit a different semantic status (cf. Hausmann, 1989). For instance, in *to pay homage* and *rendir homenaje*, the nouns are the autosemantic bases, and the collocates are *to pay* and *rendir*, i.e. verbs that are synsemantic and whose meanings are coerced by their respective bases, as well as their potential translation equivalents.

Collocations are also grammatically bounded. Some relevant authors consider that collocation is purely a lexical phenomenon which is independent of grammatical category or syntactic structure (Sinclair, 1966; Halliday and Hasan, 1976; Fontenelle, 1992). However, the grammatical dimension of collocations was soon pointed out by Mitchell (1971, p. 65), who even stated explicitly that collocations should be studied within grammatical matrices, i.e. patterns such as verb + noun, adjective + noun, verb + adverb, verb + gerund, and so forth. In fact, Hausmann's (1989) widely accepted typology of collocations is also based on grammar patterns. In the same vein, collocations have been conceived as semantic and syntactic units (Choueka, 1988), as recurring sequences that are grammatically well formed (Kjellmer, 1994), as recurrent combinations of two linguistic elements which have a syntactic relationship (Tutin, 2008) or, even, as constructions, that is, entrenched pairings of form and meaning that are

semantically predictable, contain slots to be filled by a restricted set of lexical items, and span various phrasal patterns (Corpas Pastor, 2015). In recent years syntactic structure has been increasingly established as an important defining feature, especially in the case of automatic extraction of collocations (cf. Bartsch, 2004; Seretan, 2011).

1.3. Conventionalisation and frequency

Collocation (co-)selection (restricted lexical selection), collocational sense relations and morpho-syntactic preference (semantic and syntactic boundedness) reflect knowledge of the norm conventionalised by usage. This is what distinguishes collocations from free word combinations. By way of illustration, let us consider *to pass sentence* and its Spanish counterpart *dictar sentencia*. They both mean: 'announce/state in a court of law (usually by a judge) what punishment is to be imposed upon a person convicted in a criminal proceeding'. Since both sequences are habitual ways to refer to this type of judgment and its typical context of situation, they tend to be frequently used as prefabricated chunks.

Bahns (1993, p. 253) claims that collocations spring to mind in such a way that they could be considered cognitively salient. The question is how to determine whether a given sequence is frequent or salient enough to be considered conventionalised, habitual or typical. One way is simply by counting their occurrences in a corpus, as Manning and Schütze (1999, p. 153) had suggested some time ago. A high number of occurrences would suffice to consider a collocation frequent and, therefore, deeply rooted in the language. After all, frequency of occurrence in discourse, and thus of processing, correlates with strength of entrenchment (salience); in other words, frequent repetition contributes to cognitive entrenchment in the neural network (Hoffmann, 2013, p. 315). This is in line with pre-theoretical claims of authors who stated that collocations belong to a single remembered set and that collocation is not only a statistical matter, but it has a psychological correlate (Greenbaum, 1974; Hoey, 2006[2005]).

However, raw frequencies have to be complemented by normalised figures and by association measures. In other words, a significant collocation could be considered typical and cognitively salient (entrenched), even though it might not strike as particularly frequent (see section 4 and subsections). Collocations are usually computed statistically through word distance and association strength[9]. This issue lies at the heart of Halliday's redefinition of collocation in probabilistic terms: "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of n lexical items) from an item x, the items a, b, c..." (Halliday, 1966, p. 158). It also led Jones and Sinclair (1974, p. 19) to distinguish between collocation and significant

[9] Common association measures used are mutual information (MI), chi-square (X2), phi-square (ϕ2) and log-likelihood (LR), Dice (D) and logDice, among others. For a comprehensive list of association measures see Evert (2005).

collocation: ""Collocation" is the co-occurrence of two items in a text within a specified environment. "Significant collocation" is regular collocation between items, such as that they co-occur more often than their respective frequencies and the length of text in which they appear would predict."

3. Methodological issues

In this paper we intend to establish a valid protocol to uncover varietal differences as regards collocational choices in order to promote naturalness of translated texts. Preliminary methodological issues involve the choice of corpora and corpus tools, as well as the language pairs, the nodes and the patterns for analysis.

3.1. Corpora and corpus tools

Collocational analyses usually involve very large corpora. This is even more in the case of widely spoken transnational languages. In the case of Spanish, there are some reference corpora available, like the CREA, CORPES[10] and the BYU-Davies (2002-)[11]. However, these corpora have a number of shortcomings: (i) they cannot be downloaded as a whole; (ii) their size is relatively small (especially in the case of national varieties); (iii) queries cannot be filtered per national varieties (BYU-Davies, 2002-); (iv) documents are not updated enough to be deemed representative of present day Spanish (CREA and BYU); (v) the corpora are under construction, which may compromise results (CORPES and BYU); (vi) their in-built corpus management systems are unstable and need debugging (CREA and CORPES); and (vii) the query language is rather limited (notoriously so in the case of the CREA).

One way to overcome those deficiencies is to resort to vast collections of electronic texts that are readily available as Web corpora, that is to say, giga-token corpora created by Web crawling and processing (cleaning up) with new-generation boilerplate removal and de-duplication tools. Some outstanding examples are the COW (Corpora from the Web) project (Schäfer and Bildhauer 2013), the TenTen corpus family (Suchomel and Pomikálek, 2012, Jakubíček at al., 2013) and the Aranea family of comparable Web corpora (Benko, 2014). For this paper three TenTen corpora of Spanish have been selected:

[10] The CORPES XXI – *the Reference Corpus of 21st Century Spanish* (Real Academia Española, n. d.) is a pan-Spanish general corpus of over 170 million words (1975-2014). The CORPES contains the *Reference Corpus of Contemporary Spanish* (CREA, available at http://corpus.rae.es/creanet.html) and updates. The CORPES is expected to reach over five billion words in 2018. It can be accessed at http://web.frl.es/CORPES/view/inicioExterno.view.

[11] The *Corpus del Español: 100 million words, 1200s-1900s* (BYU- Davies, 2002-) is also a pan-Spanish corpus of 100 words from the 13th to the 20th centuries. It is available online at http://www.corpusdelespanol.org. It requires registration and subscription (donation) to have full access. At present the BYU-Davies (2002-) is being updated and enlarged considerably. It is expected to reach two billion words by June 2016.

1. **esEuTenTen [2011]** – a general corpus of European Spanish (Peninsular) of two billion words (2,021,756,831 types/ 2,354,216,667 tokens). It has been crawled automatically without distinguishing among Spanish regional varieties. It will be used to analyse the Peninsular variety.
2. **esAmTenTen [2011]** – a general corpus of Latin American Spanish of seven billion words (7,475,645,291 types/ 8,640,399,540 tokens). It comprises 18 different varieties that have been identified by their national top-level domains (.ar, .es, .uy, .ve, etc.): Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay and Venezuela. 75% of the American Spanish corpus comes from Argentina, Mexico and Chile, in descending order (Suchomel and Pomikálek, 2012). The set of varieties has been crawled independently, assembled as components of the corpus and computed for similarity, as described in Kilgarriff and Renau (2013). It will be used to analyse the general American variety (as opposed to the Peninsular one) and to extract data per Latin American national varieties.
3. **esTenTen [2011]** – a pan-Spanish macrocorpus of nine billion words (9,599,765,095 types/ 10,994,616,207 tokens). It contains the esEuTenTen [2011] and the esAmTenTen [2011] (without diatopic filters). It includes 'general Spanish', understood in the sense of 'global', 'standardised' or 'unified' Spanish spoken/written across the Spanish-speaking world (cf. Paffey, 2012, p.63). It will be used to extract overall data. According to Kilgarriff and Renau (2013, p. 16), Peninsular Spanish appears to be just another variety: "The Peninsular variety shows differences with respect to other dialects, but remarkable similarities which, with the data to hand, do not distinguish it from the American varieties." The authors claim that Mexican Spanish is the variety that shows the smaller distances when compared in a one-to-one fashion to the rest; whereas Honduras, Bolivia and Paraguay seem to be the most distant varieties of all. Other interesting findings cluster Argentina and Uruguay (River Plate region) together with Chile and Peru.

All three corpora have been web-crawled with Spiderling[12], tokenised, lemmatised and part-of-speech tagged with Freeling 4.0 for tagsets (PoS classes) and lempos (PoS suffixes conjoined to lemmas)[13]. Every token (word forms or punctuation) in the corpus has assigned features or attribute values: e.g., the token *ratones* ('mice') has word (**ratones**), lemma (**ratón**), tag (**n**) and lempos (**ratón_n**).

The TenTen corpora are already pre-processed and made available through Sketch Engine[14]. This comprehensive corpus tool includes a corpus building and management system (web service) plus a corpus query tool (software) which is

[12] http://nlp.fi.muni.cz/trac/spiderling. See also Pomikalek and Suchomel (2012).
[13] https://www.sketchengine.co.uk/documentation/wiki/SkE/Help/JargonBuster.
[14] http://www.sketchengine.co.uk. For a recent description of the latest version of the tool, see Kilgariff et al. (2014).

powerful enough to process giga-token size corpora. As a matter of fact, the web service provides a large number of pre-loaded, ready-to-use corpora for more than thirty languages (designed and crawled, general language, reference and parallel, as well as some learner and historical corpora); high-level resources for eleven major languages (a tokeniser, a lemmatiser, a part-of-speech tagger and a parser); tools for creating, uploading, installing and managing users' own corpora (WebBootCat); and, of course, a corpus query tool for exploiting and analysing the corpus data: the Sketch Engine. This robust and stable software consists of three core functions:

i. **Concordance**. The concordancer (Key Word In Context or KWIC) uses an extended version of the Stuttgart corpus query language (CQL). It allows both simple and other query types for word forms, lemmas, irregular expressions, parts of speech or patterns. Searches can be further refined by context, text types and metadata, as defined by users.
ii. **Word Sketch**. This function provides a one-page summary of a word's grammatical and collocational patterns. Results are ranked according to raw frequencies or score (the salience threshold). They are also linked to the concordancer, thus users can decide on the amount of data to be shown.
iii. **Thesaurus.** The **thesaurus** offers distributional entries created on the basis of common collocation. Results are then provided by means of tables, with lemmas ranked according to score (the rate of collocational proximity), and thesaurus word clouds.

Other additional functionalities are the Good Dictionary Examples (GDEX) that helps find the best examples in a corpus, keywords and corpus comparison for any pair of same language corpora, as well as bilingual sketches and a term-finding functionality for a limited number of languages.

3.2. Nodes and patterns

Our methodology of analysis is based on lexicogrammar patterns. It comprises nodes that are quasi synonyms, appear in similar syntactic patterns and qualify as prima-face translation equivalent of semantically-functional counterparts in the target language. By restricting the choice of nodes in this way, special attention is given to collocational preferences within the same functional and semantic context. For this study we have focused on source language words that denote the act or process of reaching a resolution after consideration, resolving a question or deciding something (e.g., *decision*, *dilemma*). In consonance with our methodology, we have selected cognates that are prima-face translation equivalents (*decisión*, *dilema*), according to various bilingual dictionaries (CSD, LSD and OSD).

In at least one of their senses, the English words and their Spanish equivalents refer to the act of making a choice that involves doubt or hesitation, as illustrated by the following lexicographic definitions:

decision

1 A conclusion or resolution reached after consideration
1.1 [MASS NOUN] The action or process of deciding something or of resolving a question. (OED)

dilemma

1 A situation in which a difficult choice has to be made between two or more alternatives, especially ones that are equally undesirable
1.1. A difficult situation or problem. (OED)

decisión

1. f. Determinación, resolución que se toma o se da en una cosa dudosa. (DRAE)

dilema

2. m. Duda, disyuntiva. (DRAE)

In the enTenTen corpus[15], the English words do not appear in each other's distributional entries (size threshold= 100 items), which are generated automatically by Thesaurus. In the case of Spanish, *decisión* does not appear to be distributionally related to *dilema* as regards collocations, and vice versa (see Table 1).

DECISIÓN			**DILEMA**		
Lemma	**Score**	**Freq**	**Lemma**	**Score**	**Freq**
iniciativa	0.613	1,537,657	*interrogante*	0.363	101,639
propuesta	0.610	1,986,840	*desafío*	0.344	538,764
resolución	0.608	1,628,131	*contradicción*	0.322	194,901
acción	0.593	3,834,149	*reto*	0.319	453,125
medida	0.589	3,285,764	*paradoja*	0.311	70,018
compromiso	0.587	1,534,728	*cuestionamiento*	0.288	103,033
intervención	0.587	1,071,097	*problemática*	0.281	403,309
disposición	0.577	1,365,200	*obstáculo*	0.272	234,595
respuesta	0.574	2,273,722	*disyuntiva*	0.259	19,828

Table 1. Ten top lemmas in the distributional thesaurus (esTenTen)

In the following sections we will study the verbal collocates for *decisión* and

[15] The enTenTen [2012] corpus includes general English (American and British varieties) of eleven billion words (11,191,860,036 types/ 12,968,375,937 tokens). For the automatic detection of both varieties a classifier has been trained and applied to the web-crawled data (Jakubíček at al., 2013). In Kilgariff (2012), a keyword-based method has been used to measure distances between the two components among themselves and globally, in regard to a designed corpus (BNC) and another crawled corpus (UKWaC). The results show that enTenTen and UKWaC are the most similar two corpora, whereas enTenTen and BNC differ only slightly. It will be used to extract data for the two varieties and in general.

dilema, with special attention to their diatopic particularities in distributive terms. Our main aim is to identify the verbs which most frequently and/or typically enter into verb (V.) + noun (N.) collocations in the grammatical relation (gramrel) 'Object_of' (e.g., *adoptar una decisión*; *plantear un dilema*). In order to retrieve collocations automatically we will use the core functions of Sketch Engine (Concordance and Word Sketch), plus some of the additional functionalities. For each node, verbal collocations will be extracted and classified as regards standard Spanish and national varieties. Then, varietal differences will be established and identified (if any). And finally, we will discuss results with a view to contrastive studies and translation.

4. Automatic retrieval of V. + N. collocations

The nodes *decisión* and *dilema* appear to be widely used in the Peninsular variety, followed by Argentinian and Mexican Spanish. A comparison of the distance between nodes as regards their normalised frequencies shows that *decisión* occurs more frequently than *dilema* (194.3 difference in general Spanish vs 144.41 in Peninsular Spanish).[16] Table 2 provides raw and normalised frequency data, in general Spanish and segregated per country:

GENERAL SPANISH	2,205,792	***200.60***	69,737	***6.30***
	decisión		***dilema***	
Argentina	688,137	***79.60***	23,029	***2.70***
Bolivia	16,221	*1.90*	313	*0.04*
Chile	224,254	*26.00*	8,307	*1.00*
Colombia	177,709	*20.60*	3,471	*0.40*
Costa Rica	14,970	*1.70*	349	*0.04*
Cuba	45,623	*5.30*	2,590	*0.30*
Dominican Republic	14,539	*1.70*	364	*0.04*
Ecuador	19,130	*2.20*	412	*0.95*
El Salvador	9,424	*1.10*	260	*0.03*
Guatemala	8,250	*1.00*	254	*0.03*
Honduras	2,179	*0.30*	21	*0.0*
Mexico	384,690	***44.50***	11,728	***1.40***
Nicaragua	16,795	*1.90*	675	*0.08*
Panama	4,605	*0.50*	100	*0.01*
Paraguay	15,238	*1.80*	166	*0.02*
Peru	74,891	*8.70*	2681	*0.30*
Spain	351,691	***149.39***	11,723	***4.98***
Uruguay	48,842	*5.70*	1,328	*0.15*
Venezuela	83,928	*9.70*	1,756	*0.20*

[16] In the esAmTenTen corpus of general American Spanish, the frequency distance between *decision* and *dilema* is 207.30.

Table 2. Frequency of nodes in Spanish

Word Sketch clusters collocations with the selected nodes around five gramrels. In the case of *decisión* and *dilema*, their collocational patterns coincide with regard to types of gramrels and relative number of collocations per clusters (in descending order), with N_modifier as the most frequent type and Y_o as the least frequent one, as shown below. Examples come from the esTenTen [2011] corpus of general Spanish.

1. N_modifier (598,068 cases for *decisión* and for 20,408 for *dilema*). It comprises noun phrases composed of the substantive node and an adjective in pre- or postmodification, as in *decisión estratégica* and *dilema existencial*.
2. Object_of (588,195 cases for *decisión* and 13,477 for *dilema*). It comprises transitive and predicative V NP collocational constructions of the type *criticar una decisión* and *zanjar un dilema*.
3. Subject_of (135,580 for *decisión* and 4,982 for *dilema*). This type of collocation involves the node as subject of a transitive, intransitive or prepositional verb, eg. *competer [una decisión], estribar* [*un dilema*].
4. Modifies (305,313 for *decisión* and 1,902 for *dilema*). It involves collocation of the node as prepositional complement of another noun (N. + S.* + N.). This gramrel typically refers to the whole or the part denotated by the node (*serie de dilemas*), or participates in nominalisations of V. + N. verbal collocates, where the noun phrase nucleus tends to be a deverbal noun (*serie de dilemas*). This gramrel is typical of V. + N. nominalisations (*alcance de una decisión, abordaje de un dilema*) and various types of multiword units, such as *(estar) a la espera de una decisión*.
5. Y_o (44,103 for *decisión* and 1,493 for *dilema*). It is an 'and/or' relationship between the node and another noun (eg. *decisión y coraje* and *dilema o disyuntiva*).

As stated in section 3.3., this study will focus on V. + decisión_n and V. + dilema_n in gramrel 2. Object of, with especial attention to their diatopic distribution and peculiarities.

It is worth mentioning that some collocates which have been automatically extracted under this gramrel are, in fact, Part-of-speech (PoS) tagging or parsing errors, as illustrated below. For instance, the place adverb *aquí* (aquí_r) has been wrongly tagged as verb (V.) and classed as gramrel 2 for *dilema*, usually as part of the multiword expression *He aquí el dilema* and variants (4); the token *acerca* (acerca_r) has been wrongly tagged as acercar_v and classed as gramrel 2 (5); the token *dicha* has been wrongly tagged as dicho_j and wrongly assigned to gramrel 2 instead of gramrel 1. (N_modifier) (6).

(4) *He **aquí** el **dilema**: quedarse en casa, votar en blanco o elegir entre el muy malo y el peor.*

(5) *Todas las **decisiones acerca** de las devoluciones se tomarán a discreción*

exclusiva de Viator.

(6) *España solicitó que se anulara **dicha decisión**, por considerarla desproporcionada*.

Transitive uses are wrongly assigned to gramrel 3 in most OVS sentences (7), and vice versa (8). Similarly, transitive verbs that alternate in transitivity (causative/inchoative alternation) tend to be assigned to gramrels according to their position in the sentence, which may result in wrong assignments as well (19).

(7) *Todas las **decisiones** las **adopta** la comisión responsable del área de la Unión Europea y la gestión recae en las distintas autonomías.*

(8) *Siempre **existe** un **dilema** inevitable a la hora de elegir una revista para enviar un trabajo.*

(9) *A través de un puñado de divertidísimas ilustraciones, se nos **presentan dilemas** filosóficos tan profundos como para qué ha venido la especie humana a la tierra, qué es el miedo o si existe dios.* [sic]

PoS tagging and/or parsing errors are particularly frequent, especially in the case of intransitive verbs in OVS sentences (10); and verbs in impersonal, reflexive or 'reflexive-passive' constructions with *se* (11). All these cases have been wrongly assigned to gramrel 2 (Object_of) instead of gramrel 3 (Subject_of). Finally, another common problem is the presence of grammar, punctuation and spelling errors (12):

(10) *En el fondo de esta cuestión **reside** el **dilema** más significativo relativo a la iniciativa.*

(11) *De una rotura de menisco **surgió** la **decisión** de dejar el fútbol como practicante a los 17 años.*

(12) *Comentarios: disen* [sic] *que nadie sabe oara* [sic] *quin* [sic] *trabaja es lo que hace* [sic] *el* [sic] *vida hace* [sic] *buenos jugadore* [sic] *y otros los aprobechan* [sic] ***es** el biejo* [sic] ***dilema** del dinero concerbandolo* [sic] *hubiesen hecho mas* [sic] *de lo que le dieron por el muchacho de todos modos arriba cocoteros*.

4.1. V. + decisión_n [Object_of]

In order to obtain a manageable amont of data, searches in general Spanish have been limited to 50 items. Table 3 shows the most frequent verbal collocates for the lemma *decisión* that have been extracted by word sketch clustering, ordered by frequency rank (I-V):

I		II		III		IV		V	
1	**tomar**	11	**confirmar**	21	**anunciar**	31	**modificar**	41	**tomar+se**
2	**adoptar**	12	**justificar**	22	**comunicar**	32	**parecer**	42	**requerir**
3	**haber**	13	**ratificar**	23	**existir**	33	**faltar**	43	**fundamentar**
4	**respetar**	14	**acatar**	24	**acertar**	34	**lamentar**	44	**valorar**

5	**tener**	15	**revocar**	25	**destacar**	35	**revertir**	45	**constituir**
6	**decir**	16	**cuestionar**	26	**impugnar**	36	**defender**	46	**avalar**
7	**conocer**	17	**criticar**	27	**compartir**	37	**rechazar**	47	**revisar**
8	**esperar**	18	**aceptar**	28	**asumir**	38	**hacer**	48	**afectar**
9	**apoyar**	19	**respaldar**	29	**celebrar**	39	**dictar**	49	**implicar**
10	**apelar**	20	**determinar**	30	**aplaudir**	40	**explicar**	50	**dejar**

Table 3. The 50 most frequent verbal collocates for *decisión* (esTenTen [2011])

Notice that *tomar* plus *se* has been classed as a separate verb (tomar+se_v), *existir* and *faltar* have been assigned to the wrong gramrel; all occurrences of *constituir* and *afectar* have been classed as gramrel 2, despite their transitivity alternations; and the verb *decir* has been wrongly tagged as decir_v, as most KWIC concordance lines show collocations with dicho_j in gramrel 1 (N_modifier), as can be seen in Table 4:

que hasta hace poco remitía	***dicha***	**decisión**	a un futuro. Asif Ali Zardari
Interior) de El Cabanyal, si	***dicha***	**decisión**	fuera comunicada oficialmente al
El presente trabajo analiza	***dicha***	**decisión**	desde la óptica de la jurisprudencia
, si los hubiere.	***Dichas***	**decisiones**	serán inapelables cuando rechacen

Table 4. KWIC concordances for *dicho + decisión*

The list of top collocates changes dramatically when salience (statistical significance or association strength) is taken into account. If we look at the ten top collocates, only *tomar* (337,639/11.39) and *adoptar* (17,981/9.07) keep the first and second positions, *respetar* has advanced to the third position (8,740/7.99) and *apelar* appears now as number 4 (3,504/7.47). However, over two thirds of the very frequent verbal collocates have been relegated to lower positions: *haber* (10,749/4.09), *tener* (8,291/3.07), *decir* (4,932/4.21), *conocer* (4,490/5.60), *esperar* (4,054/6.05) and *apoyar* (3,546 /6.50); while others less frequent but significantly more salient collocates have entered this ranking: *acatar* (2,904/7.20), *revocar* (2,860/7.17), *ratificar* (2,916/6.93), *respaldar* (2,405/6.79), *cuestionar* (2,595/6.77) and *justificar* (2,948/6.76). Finally, PoS tagging and subcategorisation errors (like *decir* and *tomar*) and verbs in transitive alternations (*constituir*, *afectar*) have disappeared from the list.

Table 5 displays the 50 most salient collocates for *decisión* in the pan-corpus of general Spanish (esTenTen[2011]). Verbal collocates that occupy the same position in both frequency and significance rankings are indicated by dark cells (4%), while lighter dark cells mark collocates that have changed their position and/or frequency rank (I-V) as regards salience. 68% of collocates have changed their frequency rank (e.g., *respetar, apelar, revocar, esperar, compartir*, etc.), whereas only 20% of them have actually changed only their position in the rank (marked by an asterisk): *respetar, apelar, criticar, confirmar, aceptar, anunciar, modificar, rechazar, dictar, valorar* and *revisar*. White cells indicate

new items (22%) that do not appear in the frequency ranking, i.e. verbal collocates of relatively low frequency but high salience scores: *proferir, postergar, controverter, saludar, motivar, orientar, conocer+se, sustentar, anular, acompañar* and *notificar*.

I		II		III		IV		V	
1	**tomar**	11	**acertar**	21	**revertir**	31	**saludar**	41	**valorar***
2	**adoptar**	12	**criticar***	22	**fundamentar**	32	**motivar**	42	**anular**
3	**respetar***	13	**confirmar***	23	**tomar+se**	33	**orientar**	43	**faltar**
4	**apelar***	14	**impugnar**	24	**avalar**	34	**modificar***	44	**asumir**
5	**acatar**	15	**apoyar**	25	**proferir**	35	**conocer+se**	45	**defender**
6	**revocar**	16	**comunicar**	26	**postergar**	36	**rechazar***	46	**revisar***
7	**ratificar**	17	**aplaudir**	27	**anunciar***	37	**dictar***	47	**compartir**
8	**respaldar**	18	**esperar**	28	**recurrir**	38	**calificar**	48	**acompañar**
9	**cuestionar**	19	**lamentar**	29	**conocer**	39	**celebrar**	49	**determinar**
10	**justificar**	20	**aceptar***	30	**controvertir**	40	**sustentar**	50	**notificar**

Table 5. The 50 most salient verbal collocates for decisión (esTenTen [2011])

Column 1 (positions 1-10) and column 2 (positions 11-20) contain collocates that are both salient (among the top 20) and frequent (among the top 50), although only two of them (*tomar* and *adoptar*) keep the same position in both rankings.

It could be claimed that a list of the top 20 salient collocates (ranks 1-2) is probably a valid benchmark in the case of *decisión*, at least as general pan-Spanish is concerned. Therefore, this 20-item collocational set has been extracted from esEuTenTen (2011) and esAmTenTen (2011) for each national variety and displayed in Table 6. The goal is to compare and contrast intravarietal differences and also in regard to general Spanish. This will also ensure that collocational salience is minimally affected by direct raw frequencies of constituents. Verbal collocates have been numbered according to their positions in both salience ranks.

As indicated above (cf. section 4), the automatic extraction of collocations is not free from errors. In order to increase precision and recall, collocates consisting of a lemma plus clitics (eg. conocer+se, plantear+nos, saber+lo, tomar+se), their frequencies have been computed manually and added to the simple lemma. In the case of wrong subcategorisation, parsing or PoS tagging, the collocate has not been removed from the list as it is virtually impossible to detect all cases in the giga-token corpora. That means that the following collocate ranks should be treated with caution. By way of illustration, most examples with *acertar* (*atinar* in Costa Rica, Dominican Republic, Mexico) do not correspond to verbal transitive uses, but to noun phrases or predicative structures with the corresponding adjectives (*acertado, atinado*).

V. + *decisión*_n (Object_of)		
	I	**II**
GENERAL	1) tomar, 2) adoptar, 3) respetar, 4)	11) acertar, 12) criticar, 13) confirmar,

SPANISH	apelar, 5) acatar, 6) revocar, 7) ratificar, 8) respaldar, 9) cuestionar, 10) justificar	14) impugnar, 15) apoyar, 16) comunicar, 17) aplaudir, 18) esperar, 19) lamentar, 20) aceptar
Argentina	1) tomar, 2) adoptar, 3) apelar, 4) ratificar, 5) respetar, 6) cuestionar, 7) acatar, 8) revocar, 9) justificar, 10) comunicar	11) criticar, 12) respaldar, 13) postergar, 14) fundamentar, 15) conocer, 16) avalar, 17) rever, 18) confirmar, 19) acompañar, 20) aguardar
Bolivia	1) tomar, 2) apelar, 3) acatar, 4) reconsiderar, 5) saludar, 6) lamentar, 7) respetar, 8) acertar, 9) aplaudir, 10) ponderar	11) ratificar, 12) adoptar, 13) revocar, 14) respaldar, 15) postergar, 16) impugnar, 17) asumir, 18) consensuar, 19) revertir, 20) criticar
Chile	1) tomar, 2) adoptar, 3) acatar, 4) revertir, 5) lamentar, 6) postergar, 7) revocar, 8) respaldar, 9) acertar, 10) respetar	11) valorar, 12) criticar, 13) impugnar, 14) fundamentar, 15) comunicar, 16) apelar, 17) ratificar, 18) cuestionar, 19) justificar, 20) orientar
Colombia	1) ratificar, 2) cuestionar, 3) justificar, 4) orientar, 5) proferir, 6) revocar, 7) impugnar, 8) tomar, 9) adoptar, 10) controvertir	11) apelar, 12) acatar, 13) confirmar, 14) notificar, 15) sustentar, 16) fundamentar, 17) acertar, 18) aplazar, 19) ejecutoriar, 20) motivar
Costa Rica	1) tomar, 2) apelar, 3) reconsiderar, 4) atinar, 5) acertar, 6) fundamentar, 7) impugnar, 8) adoptar, 9) posponer, 10) respaldar	11) respetar, 12) aplaudir, 13) revocar, 14) sustentar, 15) notificar, 16) postergar, 17) anular, 18) comunicar, 19) avalar, 20) lamentar
Cuba	1) tomar, 2) patentizar, 3) adoptar, 4) acertar, 5) apelar, 6) controvertir, 7) ratificar, 8) acatar, 9) reafirmar, 10) aplaudir	11) condenar, 12) reconsiderar, 13) revocar, 14) repudiar, 15) respaldar, 16) vetar, 17) saludar, 18) justar, 19) doblegar, 20) apresurar
Dominican Republic	1) desacatar, 2) saludar, 3) apelar, 4) atinar, 5) tomar, 6) adoptar, 7) revocar, 8) acatar, 9) ponderar, 10) acertar	11) invalidar, 12) acoger, 13) respaldar, 14) lamentar, 15) respetar, 16) criticar, 17) recurrir, 18) impugnar, 19) protestar, 20) aplaudir
Ecuador	1) tomar, 2) apelar, 3) acertar, 4) aplaudir, 5) rever, 6) adoptar, 7) conocer, 8) aplazar, 9) saludar, 10) respetar	11) respaldar, 12) reconsiderar, 13) acatar, 14) ratificar, 15) oficializar, 16) democratizar, 17) revocar, 18) impugnar, 19) lamentar, 20) felicitar
El Salvador	1) amarrar, 2) tomar, 3) acertar, 4) aplaudir, 5) apelar, 6) impugnar, 7) revocar, 8) adoptar, 9) acatar, 10) criticar	11) respetar, 12) lamentar, 13) respaldar, 14) cuestionar, 15) notificar, 16) basar, 17) fundamentar, 18) revertir, 19) avalar, 20) guiar
Guatemala	1)tomar, 2) impugnar, 3) apelar, 4) acertar, 5) revocar, 6) aplaudir, 7) conocer, 8) acatar, 9) respetar, 10) revertir	11) saludar, 12) notificar, 13) respaldar, 14) criticar, 15) avalar, 16) adoptar, 17) lamentar, 18) conllevar, 19) cuestionar, 20) aceptar
Honduras	1) alabar, 2) tomar, 3) apelar, 4) aplaudir, 5) saludar, 6) calificar, 7) lamentar, 8) revertir, 9) respaldar, 10) cuestionar	11) avalar
Mexico	1) tomar, 2) respetar, 3) acatar, 4) aplaudir, 5) apelar, 6) acertar, 7) respaldar, 8) impugnar, 9) adoptar,	11) atinar, 12) revocar, 13) avalar, 14) criticar, 15) posponer, 16) ratificar, 17) apoyar, 18) revertir, 19) cuestionar, 20)

	10) lamentar	sustentar
Nicaragua	1) tomar, 2) apelar , 3) aplaudir, 4) acatar, 5) acertar, 6) reconsiderar, 7) conocer, 8) revocar, 9) revertir, 10) lamentar	11) respaldar, 12) precipitar, 13) respetar, 14) impugnar,15) adoptar, 16) elogiar, 17) protestar, 18) criticar, 19) ratificar, 20) cuestionar
Panama	1) apelar, 2) tomar, 3) conocer, 4) reconsiderar, 5) acatar, 6) acertar, 7) impugnar, 8) revocar, 9) aplaudir, 10) adoptar	11) respaldar, 12) lamentar, 13) notificar, 14) criticar, 15) respetar, 16) ratificar, 17) recurrir
Paraguay	1) rever, 2) impugnar, 3) revocar, 4) acatar, 5) acertar, 6) apelar, 7) tomar, 8) aguardar, 9) adoptar, 11) repudiar	11) fundamentar, 12) basar, 13) sustentar, 14) dilatar, 15) cuestionar, 16) recaer, 17) aplaudir, 18) torcer, 19) comunicar, 20) rectificar
Peru	1) tomar, 2) saludar, 3) apelar, 4) acertar, 5) adoptar, 6) acatar, 7) respaldar, 8) respetar, 9) cuestionar, 10) impugnar	11) revocar, 12) reconsiderar, 13) aplaudir, 14) lamentar, 15) desatinar, 16) ratificar, 17) influenciar, 18) controvertir, 19) postergar, 20) sustentar
Spain	1) tomar 2) adoptar, 3) respetar, 4) recurrir, 5) acatar, 6) justificar, 7) criticar, 8) aplaudir, 9) modificar, 10) derogar	11) acertar, 12) comunicar, 13) apoyar, 14) apelar, 15) aplazar, 16) respaldar, 17) aceptar, 18) ratificar, 19) revocar,20) anular
Uruguay	1) tomar, 2) rever, 3) adoptar, 4) acatar, 5) apelar, 6) reconsiderar, 7) postergar, 8) acertar, 9) revocar, 10) desacatar	11) respaldar, 12) respetar, 13) aguardar, 14) fundamentar, 15) comunicar, 16) cuestionar, 17) aplaudir, 18) criticar, 19) ratificar, 20) precipitar
Venezuela	1) acatar, 2) tomar, 3) apelar, 4) revocar, 5) dictar, 6) impugnar, 7) anular, 8) acertar, 9) fundamentar, 10) desacatar	11) aplaudir, 12) proferir, 13) suscribir, 14) referir, 15) reconsiderar, 16) respaldar, 17) recaer, 18) adoptar, 19) ratificar, 20) saludar

Table 6. Diatopic varieties of salient verbal collocates for *decisión*

A cursory look at Table 6 shows that not all national varieties share a similar degree of lexical richness as to their preferred verbal collocates for *decisión*. All varieties list ten significant verbs in Rank I, though Rank II contains a limited number of collocates in the case of Honduras (11) and Panama (17). Further distributional differences apply with regard to the salience of a given verbal collocate in general Spanish and the different national varieties. For instance, *respaldar una decisión* appears in rank I (position no. 8) in general Spanish. The collocation is documented in all national varieties. However, it is not significant for Paraguay and Colombia (salient score 0.0). In the rest of the countries (17), *respaldar una decisión* is a salient collocation which appears in Rank I of five national varieties (29.41%), in descending order: Mexico and Peru (position 7), Chile (position 8), Honduras (position 9), Costa Rica (position 10); and in Rank II of the rest (70.58%): Ecuador, Nicaragua, Panama and Uruguay (position 11), Argentina (position 12), Dominican Republic, El Salvador, Guatemala (position 13), Bolivia (position 14), Cuba, Spain and Venezuela (position 16).

Diatopically restricted collocations are to be found in cases where the collocates are verbs belonging to a particular language variety: e.g., *rever* (‘review, retry’) is typical of Argentina, Ecuador, Paraguay and Uruguay; or *desacatar* (‘disobey’), which is used in the Dominican Republic, Uruguay and Venezuela. Other diatopic peculiarities can be observed in common verbs in general Spanish which have developed specialised, coerced collocational meanings in certain national varieties due to semantic or pragmatic mismatches. For example, the verb *proferir* (‘utter’, ‘hurl’) usually combines with nouns denoting insults, words, sounds and curses (see Table 7), though in Colombia and Venezuela it usually collocates with *decisión* in the legal sense of ‘pass sentence or judgement’ (see Table 8).

quieren sancionar a los jugadores que	**profieran**	***insultos*** blasfemos. ESPN, 15 de febrero
las alternativas con el vecino y vuelve a	**proferir**	alguna ***exclamación*** sofocada por la voz
a su designio, se extiende, apenas	**proferida**	, hacia otras ***palabras*** , formando una
1. Que las personas que	**profieran**	por Internet una ***expresión*** que genere

Table 7. KWIC concordances for *proferir + N.**(esAmTenTen)

consolidado antes de ***proferirse*** la	**decisión**	del Consejo de Estado ya
que así lo declare, deberá ***proferirse*** la	**decisión**	motivada que resuelva el recurso.
su defensa. Sin embargo al ***proferirse*** la	**decisión**	en segunda instancia se varió esta
> ...ANULA las ***decisiones***	**proferidas**	por el Tribunal Primero de Primera Instancia

Table 8. KWIC concordances for *proferir + decisión* (Colombia and Venezuela).

There is also restriction of the combinatory properties of certain verbs, as in the case of *patentizar* (‘illustrate’, ‘demonstrate’) in Cuban Spanish; in collocation with *decisión*, the verb has undergone semantic and pragmatic specialisation in the sense of ‘making clear or evident a personal or political choice, that is also sanctioned positively by the community’ (see Table 9).

por el Primero de ***Mayo*** , donde sus hijos	**patentizarán**	su apoyo a las ***decisiones*** adoptadas por
profesionales ***guantanameros*** , quienes	**patentizaron**	la ***decisión*** de mantenerse en la primera
contr un ***niño*** indefenso, los villaclareños	**patentizan**	la ***decisión*** del pueblo cubano de
de Jóvenes Comunista s (***UJC***) y se	**patentizó**	la ***decisión*** de proseguir la lucha por el

Table 9. KWIC concordances for *patentizar + decisión* (Cuba)

Finally, a comparative study of *V.* + decisión_n in general Spanish and the three national varieties with the largest corpus size (Argentina, Mexico and Spain) shows a clear tendency for diatopic preferences as to the choice of verbal collocates. As Table 10 illustrates, the total number of significant verbal collocates in the three national varieties is 37, of which 31.45% are diatopically restricted and 68.55% coincide (at least partially) with general Spanish. As

regards the choice of verbal collocates, the Mexican variety seems to be closer to general Spanish (60% of shared verbs), in comparison with Argentinian and Peninsular Spanish, which both share 45% of their verbal collocates. The three national varieties also exhibit different degrees of cross-varietal similarity: of the whole set of collocational verbs (37), Mexico-Spain coincide in the highest percentage (32.43%), followed by Mexico-Argentina (29.72%) and Argentina-Spain (24.32 %), the most distant varieties. Diatopic preferences are also visible in the choice of verbs for particular semantic and functional values. For instance, the act of deferring to take a decision is typically conveyed by different verbs in the three national varieties under comparison: *postergar* (Argentina), *posponer* (Mexico) and *aplazar* (Spain) *una decisión*. Other examples are verbal collocates used to express accepting or observing of a decision (with nuances), such as *adoptar/respetar/acatar una decisión* (general Spanish, Argentina, Mexico and Spain), *aplaudir una decisión* (Mexico, Spain) and *aceptar una decisión* (Spain); as well as salient verbal collocates for the expression of supporting a decision: *respaldar una decisión* (general Spanish, Argentina, Mexico and Spain), *apoyar una decisión* (general Spanish, Mexico and Spain) and *avalar una decisión* (Argentina, Mexico).

DECISIÓN	**GEN. SPANISH**	**ARGENTINA**	**MEXICO**	**SPAIN**
tomar	x	x	x	x
adoptar	x	x	x	x
respetar	x		x	x
apelar	x	x	x	x
acatar	x	x	x	x
revocar	x	x	x	x
ratificar	x	x	x	x
respaldar	x	x	x	x
cuestionar	x	x	x	
justificar	x	x		x
acertar	x		x	x
criticar	x	x	x	x
confirmar	x	x		
impugnar	x		x	
apoyar	x		x	x
comunicar	x	x		x
aplaudir	x		x	x
esperar	x			
lamentar	x		x	
aceptar	x			x
ARGENTINA				
postergar		x		
fundamentar		x		
conocer		x		
avalar		x	x	
rever		x		

acompañar		x		
aguardar		x		
MEXICO				
atinar			x	
posponer			x	
revertir			x	
sustentar			x	
SPAIN				
recurrir				x
modificar				x
derogar				x
aplazar				x
anular				x

Table 10. Verbal collocates for *decisión* (general Spanish, Argentina, Mexico and Spain)

4.2. *V.* + dilema_n [Object_of]

The methodology described in section 4.1. was be applied for the analysis of the verbal collocates for the second node selected. First, the fifty most frequent verbal collocates for the noun *dilema* were retrieved automatically (Table 11). In order to minimise errors and increase precision and recall, lemmas plus clitics (eg. plantear+se, planter+nos) have been computed and added manually (eg. *plantear*), and adjectives with the wrong lempos (eg. *intrincado* as intrincar_v or *cacareado* as cacarear_v) were eliminated from the list. In addition, intransitive verbs like *surgir*, *residir* or *florecer* were deleted from the list as all KWIC concordances showed gramrel 3 (Subject_Of) or gramrel 1 (N_Modifier). However, some PoS tagging and parsing errors remain, as it is not possible to eliminate all of them automatically. For example, the lemma *abrir* has been automatically assigned the tag .V* and classed as gramrel 2 by Sketch Engine, even though the corresponding word forms are sometimes adjectives (estar/seguir + abierto_j) or enter into gramrel 1 and other types of attributive syntactic relationships (dejar + abierto_j), as can be seen below (13-14).

(13) *Está **abierto** el **dilema** entre un acuerdo "light" y un acuerdo "ambicioso".*

(14) *En realidad, estos modelos dejan **abierto** el **dilema** sobre el incremento de la viabilidad real y topográfica.*

I		II		III		IV		V	
1	**plantear**	11	**afrontar**	21	**evitar**	31	**reflejar**	41	**confrontar**
2	**resolver**	12	**abordar**	22	**comprender**	32	**encerrar**	42	**identificar**
3	**tener**	13	**representar**	23	**abrir**	33	**proponer**	43	**tratar**
4	**enfrentar**	14	**constituir**	24	**provocar**	34	**exponer**	44	**formular**
5	**presentar**	15	**suponer**	25	**encarar**	35	**expresar**	45	**describir**
6	**existir**	16	**venir**	26	**complicar**	36	**dilucidar**	46	**aclarar**
7	**generar**	17	**vivir**	27	**implicar**	37	**suscitar**	47	**imponer**

8	**crear**	18	**analizar**	28	**discutir**	38	**resumir**	48	**causar**
9	**superar**	19	**aparecer**	29	**romper**	39	**empezar**	49	**atravesar**
10	**solucionar**	20	**entender**	30	**ilustrar**	40	**explorar**	50	**eludir**

Table 11. The 50 most frequent verbal collocates for *dilema* (esTenTen [2011])

As it could easily be predicted, salience (association strength) alters significantly the list of verbs that typically co-occur with the word *dilema* (see Table 12). Thus, the number of salient verbal collocates has been reduced to 48. Besides, only seventeen verbs (34%) of the frequency ranking are also present in the salience ranking: *plantear* and *resolver* remain in the first two positions, whereas the other fifteen have changed ranks, with the notable exception of *enfrentar* which has only changed position within Rank I. The rest (33 verbs, 66%) have entered the salience rank anew. This has led to dramatic changes in the positions of frequent verbs in the corpus once filtered by statistical significance: e.g. *confrontar* changes from position 41 in Rank V to position 6 in Rank I; *tener* (position 3 Rank I) does not make it to the salience ranking; and *zanjar*, a relatively low-frequency verb not listed in the previous ranking, occupies the fifth position in the salience ranking.

I		II		III		IV		V	
1	**plantear**	11	**ilustrar**	21	**abordar**	31	**despejar**	41	**esquivar**
2	**resolver**	12	**suscitar**	22	**desentrañar**	32	**explorar**	42	**debatir**
3	**enfrentar ***	13	**subyacer**	23	**resumir**	33	**disolver**	43	**examinar**
4	**dilucidar**	14	**encarar**	24	**encarnar**	34	**entrañar**	44	**trascender**
5	**zanjar**	15	**replantear**	25	**ejemplificar**	35	**sortear**	45	**persistir**
6	**confrontar**	16	**intrincar**	26	**superar**	36	**presentar**	46	**emerger**
7	**afrontar**	17	**aparejar**	27	**retratar**	37	**discutir**	47	**recrear**
8	**solucionar**	18	**eludir**	28	**agudizar**	38	**solventar**	48	**representar**
9	**encerrar**	19	**descifrar**	29	**esbozar**	39	**enfocar**	49	
10	**dirimir**	20	**complicar**	30	**clarificar**	40	**esclarecer**	50	

Table 12. The 50 most salient verbal collocates for *dilema* (esTenTen [2011])

Rank I contains collocates that are both frequent and salient (80%, the exceptions being *zanjar* and *dirimir*). Rank II contains just 30%. This means that, unlike the previous case, a benchmark of 10 (Rank I of salient collocates) could be valid in the case of *dilema*. A plausible explanation could be the lower frequency of occurrence of this lemma in the corpus, as compared with *decisión*, which occurs almost thirty two times more per million (cf. Table 2). In any case, Rank I and Rank II have been taken into account in order to assess collocational differences and preferences among national varieties and with regards to general Spanish (see Table 13).

V. + dilema_n (Object_of)		
	Rank I	**Rank II**
GENERAL	1) plantear, 2) resolver , 3)	11) acertar, 12) criticar , 13) confirmar,

SPANISH	enfrentar, 4) dilucidar, 5) zanjar, 6) confrontar, 7) afrontar, 8) solucionar, 9) encerrar, 10) dirimir	14) impugnar, 15) apoyar, 16) comunicar , 17) aplaudir, 18) esperar, 19) lamentar, 20) aceptar
Argentina	1) plantear, 2) resolver, 3) dilucidar, 4) enfrentar, 5) aparejar, 6) encerrar , 7) solucionar, 8) afrontar, 9) encarar, 10) suscitar	11) confrontar, 12) esquivar, 13) eludir, 14) resumir, 15) encarnar, 16) complicar
Bolivia	1) contornar, 2) problematizar, 3) plantear, 4) desentrañar, 5) rastrear, 6) confrontar	
Chile	1) zanjar, 2) resolver, 3) plantear, 4) enfrentar, 5) retratar, 6) encerrar, 7) afrontar, 8) eludir , 9) solucionar, 10) abordar	11) encarar , 12) surgir , 13) superar
Colombia	1) ilustrar, 2) dirimir, 3) plantear, 4) resolver, 5) confrontar, 6) enfrentar, 7) afrontar, 8) recrear, 9) encarar , 10) solucionar	11) surgir, 12) abordar
Costa Rica	1)resolver, 2) añadir	
Cuba	1) descifrar, 2) plantear , 3) solucionar, 4) resolver, 5) afrontar, 6) enfrentar, 7) resumir	
Dominican Republic	1) plantear, 2) discernir, 3) dilucidar	
Ecuador	1) replantear	
El Salvador	1) ahorrar, 2) resultar, 3) interrogar, 4) clarificar, 5) dirimir	
Guatemala		
Honduras		
Mexico	1) plantear, 2) enfrentar , 3) resolver , 4) confrontar, 5) dirimir, 6) ilustrar, 7) suscitar, 8) encerrar, 9) afrontar, 10) solucionar	11) encarar , 12) abordar , 13) explorar, 14) trascender , 15) superar, 16) complicar, 17) romper, 18) discutir, 19) conllevar, 20) examinar
Nicaragua	1) escrutar, 2) desvelar, 3) desenmascarar, 4) relucir , 5) saber, 6) resolver, 7) enfrentar	
Panama	1) imaginar, 2) confrontar	
Paraguay	1) plantear, 2) resolver, 3) enfrentar, 4) aclarar, 5) asumir, 6) presentar, 7) conocer , 8) poner , 9) deber , 10) tener	
Peru	1) plantear , 2) enfrentar, 3) resolver, 4) afrontar	
Spain	1) plantear, 2) resolver, 3) enfrentar, 4) solucionar, 5) complicar, 6) esclarecer, 7) despejar, 8) suscitar, 9) afrontar, 10) solventar	11) abordar, 12) ilustrar, 13) explorar, 14) discutir, 15) presentar, 16) resumir, 17) reflejar, 18) superar, 19) aclarar, 20) suponer
Uruguay	1) dilucidar, 2) plantear, 3) enfrentar, 4) encerrar, 5) resolver,	

	6) resumir	
Venezuela	1) plantear, 2) dilucidar, 3) confrontar, 4) resolver, 5) encarnar, 6) enfrentar, 7) encarnar, 8) afrontar	

Table 13. Diatopic varieties of salient verbal collocates for *dilema*

The first noticeable fact is the differences in collocational richness across national varieties. Only Mexico and Spain list 20 salient verbal collocates for *dilema* (as in the case of general Spanish). In the rest of the countries the number of different verbal collocates varies substantially: 16 (Argentina), 13 (Chile), 12 (Colombia), 10 (Paraguay), 8 (Venezuela), 7 (Cuba and Nicaragua), 6 (Bolivia and Uruguay), 5 (El Salvador), 3 (Dominican Republic), 3 (Peru), 2 (Costa Rica and Panama), 1 (Ecuador and Guatemala), and 0 (Honduras). For some national varieties there was insufficient data available to retrieve word sketches (Panama and Paraguay), so the additional functionality Collocations has been used instead in order to extract candidate collocates. For Honduran and Guatemalan varieties it has not been possible to retrieve automatically salient collocations through Word Sketch nor Collocations. However, general Spanish salient collocates can be found in the in both subcorpora, especially those in Rank I, as evidenced for Honduras (15-16) and Guatemala (17-18).

(15) *El estamento hondureño **se enfrentó** a un **dilema**: unanimidad casi absoluta entre las instituciones del Estado y la clase política en que Zelaya había abusado de sus poderes en violación de la Constitución, pero con cierta ambigüedad sobre qué hacer al respecto*.

(16) *Ahora bien, tratando de **resolver** para mí misma, el segundo **dilema planteado**, diré que, en mi experiencia a lo largo de mis 36 años de vida, he conocido personas a quienes su oficio ha etiquetado como "maestros constructores o de obra".*

(17) *Declarado santo por la Iglesia católica, Tomás Moro es un ejemplo de vida para quienes deben **afrontar dilemas** entre ética y política.*

(18) *A Desde chiquitos, nuestros hijos nos **plantean dilemas** que tienen que ver con la ética.*

Even though there are fewer verbal collocates for *dilema* than for *discusión*, and there is no verbs diatopically restricted (cf. *rever* in 4.1), diatopic variation is manifested not only in the varying degrees of collocational richness but also in the collocational preferences observed for each national variety. For instance, in general Spanish the act of facing a dilemma would be translated as *afrontar/confrontar/enfrentar un dilema*. National varieties would select one or several verbs from the former list or, else, the synonym *encarar* (Chile, Mexico and Colombia), though some countries show clear collocational preferences. Mexico and Colombia tend to use the four verbal alternatives (*afrontar/confrontar/enfrentar/encarar)*, whereas the rest of the countries select

only some of them: *afrontar/confrontar/enfrentar* (Venezuela and Argentina), *afrontar/encarar, enfrentar* (Chile), *afrontar/enfrentar* (Cuba, Peru and Spain), *confrontar* (Bolivia, Panama) and *enfrentar* (Nicaragua, Paraguay and Uruguay).

As in seen in section 4.1., the comparative study of V. + dilema_n in general Spanish and the three national varieties with the largest corpus size (Argentina, Mexico and Spain) illustrates further the existence of diatopic preferences as regards verbal collocates (see Table 14). The number of different verbal collocates present in the three varieties altogether (45) is higher than for *decisión* (37). Less than half (44.44%) belong to general Spanish, whereas 55.55% of the verbs are typical of the national varieties, with Mexico being the richest (10), followed by Argentina (8) and Spain (7). Mexican and Argentinian varieties seem to be closer to general Spanish (40% of shared verbs), while Peninsular Spanish appears to be more distant (only 20% of shared verbs). In terms of cross-varietal similarity, again Mexican and Spanish varieties appear to be in close proximity (60.00 % of shared verbs), followed by Mexican and Argentinian varieties (37.5%), and Argentinian and Spanish varieties (12.5%). This is in line with the findings in section 4.1. and with the claims about distance between Spanish national varieties by Kilgarriff and Renau (2013).

DECISIÓN	**GEN. SPANISH**	**ARGENTINA**	**MEXICO**	**SPAIN**
plantear	x	x	x	x
resolver	x	x	x	x
enfrentar	x	x	x	x
dilucidar	x	x		
zanjar	x			
confrontar	x	x	x	
afrontar	x	x	x	x
solucionar	x	x	x	x
encerrar	x	x	x	
dirimir	x		x	
acertar	x			
criticar	x			
confirmar	x			
impugnar	x			
apoyar	x			
comunicar	x			
aplaudir	x			
esperar	x			
lamentar	x			
aceptar	x			
ARGENTINA				
aparejar		x		
encarar		x	x	
suscitar		x	x	x
esquivar		x		
eludir		x		

resumir		x		x
encarnar		x		
complicar		x	x	x
MEXICO				
ilustrar			x	x
abordar			x	x
explorar			x	x
trascender			x	
superar			x	x
romper			x	
discutir			x	x
conllevar			x	
complicar			x	x
examinar			x	
SPAIN				
esclarecer				x
despejar				x
solventar				x
presentar				x
reflejar				x
aclarar				x
suponer				x

Table 14. Verbal collocates for *dilema* (general Spanish, Argentina, Mexico and Spain)

Further examples of cross-varietal distances and with regard to general Spanish can be found in the selection of verbs which convey posing or solving a dilemma. In the first case, general Spanish verbal collocates are *plantear* and *encerrar*. The list is enlarged with the following diatopically restricted verbal collocates: *aparejar*, *suscitar*, *encarnar* (Argentina); *conllevar* (Mexico); *reflejar* and *presentar* (Spain). In the second case, the list of general Spanish verbs which are salient in collocation with *dilema* are *resolver*, *dilucidar*, *zanjar*, *solucionar*, *dirimir* and *acertar*. Argentinian Spanish favours the typical collocates of general Spanish. However, Mexican Spanish adds *superar* and *romper*, whereas Peninsular Spanish appears the richest in peculiar significant collocates: *esclarecer, despejar, solventar* and *aclarar*.

5. Conclusion

Combinatory restrictions reflect a language's idiosyncrasy at all levels, including diasystematic variation. Collocations are conventionalised word combinations which are frequent and/or salient, and exhibit restricted lexical selection, morpho-syntactic preference and semantic boundedness. In translation, collocations pose problems mainly in production, especially in the choice of the right collocates. In the case of transnational languages, collocability is actually governed by the peculiarities of the given target language variety.

However, in translation very little attention has been paid to this fact. Bilingual dictionaries provide a poor coverage of collocations, insufficient microstructural information and simplified, prototypical translation equivalents. By way of illustration, let us consider the case of *to postpone a decision* and *to pose a dilemma*. The two collocations are not included in CSD, LSD nor OSD. Translated texts also show a tendency towards simplification, as well as equivalence inaccuracies and variety bias. In the case of *decision*, the translations retrieved by Linguee show the following equivalent verbal collocates for expressing deferring a decision-taking act: *aplazar* (17 occurrences, 58.62 %), *posponer* (10/ 34, 48 %), *postergar* (1, 0.68%) and *retrasar*[17] (1, 0.68 %). These results are indicative of a tendency towards Peninsular Spanish to the detriment of the Argentinan and Mexican varieties. As to the English collocation *to pose a dilemma*, Linguee retrieves *plantear* (10 occurrences, 90.90%) plus a modulation with *ser − ser* sth. *un dilema para* sth./sb.− (1, 9.09%). The results show a clear preference for simplification and normalisation, in line with general Spanish: the most frequent/salient verbal collocate (*plantear*) is preferred over the second one (*encerrar*), whereas diatopically-restricted collocates have been avoided. Should this be a generalised tendency, the translation of collocations could play a key role in the (machine learning) identification of translationese and of translation universals such as simplification and normalisation (cf. Corpas Pastor, 2008; Ilisei et al., 2010).

Spanish is not a monolithic language. It consists of unified, general Spanish and national language varieties. Processing of giga-token corpora can help uncover relevant features of general Spanish as opposed to the diatopically restricted peculiarities of national varieties. In this respect, collocations and large corpora appear to be crucial. The results of our study support the claim by Kilgarriff and Renau (2013) about Mexican Spanish being the variety that shows the smaller distances when compared in a one-to-one fashion to the rest. They also evidence a closer collocational proximity between the Mexican and the Peninsular varieties, as well as different degrees of cross-varietal collocational similarity. Diatopic preferences are visible in the varying collocational richness of national varieties, in the differences as regards their frequency and salience rankings, and in their idiosyncratic selection of collocates to convey particular semantic and functional values.

Finally, the results in this study should be treated with caution, as they could have been affected by the corpora and the NLP tools used. Corpus-based automatic retrieval of collocations should be further refined in order to reduce the presence of grammar, punctuation and spelling mistakes, as well as PoS tagging and/or parsing errors. In the first case, corpus preparation and processing (document selection and cleaning) should be improved; in the second case, more robust parsing and annotation systems should be in place, especially for OVS

[17] The results also evidence the use of non salient collocations, such as *retrasar una decisión*.

languages. A possible way forward could be to apply shallow semantic parsing (semantic role labelling). For example, this would help to disambiguate between gramrels 2 (Object_Of) and 3 (Subject_Of), a common problem for extracting verbal collocations. It would be interesting to extend this methodology of analysis to more words, especially synonyms and words belonging to different frequency ranks, as well as to other collocational gramrels (or patterns), to other transnational languages, regional varieties within the same language or national varieties, and to translated versus non-translated texts.

Acknowledgements

The research presented in this paper has been partially carried out in the framework of research projects Expert (317471-FP7-PEOPLE-2012-ITN) and Inteliterm (FFI2012‑38881).

References

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT Journal, 1*(47), 56–63.

Bartsch, S. (2004). *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Narr.

Barnbrook, G., Mason, O. & Krishnamurthy, R. (2013). *Collocation: applications and implications*. Basingstoke: Palgrave Macmillan.

Benko, V. (2014). Aranea: yet another family of (comparable) web corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, speech and dialogue. 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655* (pp. 257-264). Springer International Publishing Switzerland.

Biber, D. (2011). Corpus linguistics and the study of literature. Back to the future?. *Scientific Study of Literature, 1*(1), 15-23.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Choueka, Y. (1988). Looking for needles in a haystack or: locating interesting expressions in large textual databases. In *Proceedings of the International Conference on user-oriented content-based text and image handling* (pp. 609–623). Cambridge, Massachusetts, USA.

Corpas Pastor, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

Corpas Pastor, G. (2001). Corrientes actuales de la investigación fraseológica en Europa. *Euskera, 46* (1), 21-49.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

Corpas Pastor. G. (2015). Register-specific Collocational Constructions in English and Spanish: a Usage-based Approach. *Journal of Social Sciences*. Retreieved from http://thescipub.com/PDF/ofsp.9994.pdf

Cowie, A. P. (1981). The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics, 2* (3), 223-235.

Davies, M. (2002-). *Corpus del Español: 100 million words, 1200s-1900s*. http://www.corpusdelespanol.org.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. London: Oxford University Press.

Firth, J. R. (1968). Linguistic analysis as a study of meaning. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952-59* (pp. 12-26). London/Harlow: Longmans.

Fontenelle, T. (1992). Collocation acquisition from a corpus or from a dictionary: a comparison. In *Proceedings I-II. Papers submitted to the 5th EURALEX international congress on lexicography on Tampere*. 221-228.

Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr.

Greenbaum, S. (1974). Some Verb-intensifier Collocations in American and British English. *American Speech,49* (1-2), 79-89.

Halliday, M.A.K. 1966. Lexis as a linguistic level. In C. Bazell, J. C. Catford, M.A.K. Halliday & R. H. Robins (Eds.), *In memory of John Firth* (pp. 148-162). London: Longman.

Halliday, M.A.K. & R. Hasan (1976). *Cohesion in English*. Longman: London.

Hardy, D. E. (2004). Collocational analysis as a stylistic discovery procedure: the case of Flannery O'Connor's Eyes. *Style, 38* (4), 410-27.

Hausmann, F. J. (1989). Le dictionnaire de collocations. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (Eds.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationals Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie* (pp. 1000-1019). Vol. I. Berlin/New York: Walter de Gruyter.

Hilper, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory, 2*(2), 243-57.

Hoover, D. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing, 18* (3), 261-286.

Hoey, M. (2006 [2005]). *Lexical priming*. London/New York: Routledge.

Hoffmann, T. (2013). Abstract Phrasal and Clausal Constructions. In T. Hoffmann & G. Trousdale (Eds.). *The Oxford handbook of construction grammar* (pp. 307-328). Oxford: Oxford University Press.

Ilisei, I. Inkpen, D., Corpas Pastor, G. & R. Mitkov, R. 2010. Identification of translationese: A machine learning approach. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 503-511). Berlin/Heildelberg: Springer.

Jakubíček, M.; Kilgarriff, A.; Kovář, V.; Rychlý, P. & V. Suchomel. (2013). The TenTen corpus family In *Proceedings of the 7th international corpus linguistics conference CL 2013* (125-127). United Kingdom, July 2013.

Jones, S. & J. M. Sinclair. (1974). English Lexical Collocations. A Study in Computational Linguistics. *Cahiers de Lexicology, 24*, 15-61.

Kilgarriff, A. (2012). Getting to know your corpus. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Proceedings of the 15th international conference on text, speech and dialogue (TSD)* (pp. 3-15). Czech Republic, September 2012.

Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J; Rychlý, P.; & V. Suchomel (2014). The Sketch Engine: ten years on. *Lexicography: Journal of ASIALEX*, *1* (1), 7-36.

Kilgarriff, A. & I. Renau (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia Social and Behavioral Sciences, 95*, 12-19.

Kjellmer, G. (1994). *A dictionary of English collocations*. Oxford: Clarendon Press.

Koike, K. (2001). *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico.* Alcalá de Henares: Universidad de Alcalá / Takushoku University.

Manning, C. D. & H. Schütze. (1999). *Foundations of statistical natural language processing*. The MIT Press: Cambridge, Massachusetts.

Mitchell, T. F. (1971). Linguistic 'goings on': collocations and other lexical matters arising on the syntactic record. *Archivum Linguisticum*, *2*, 35–69.

Molero, A. (2003). *El español de España y el español de América*. *Vocabulario comparado*. Madrid: Ediciones SM.

Paffey, D. (2012). *Language ideologies and the globalization of 'standard' Spanish*. (Advances in Sociolinguistics). Bloomsbury Publishing: London/New York.

Quirk, R; Greenbaum, S.; Leech, G.; & J. Svartvik (1989). *A comprehensive grammar of the English language*. London: Edward Arnold.

Real Academia Española (n.d.). Banco de datos (CORPES XXI) [on line]. Corpus del español del siglo XXI. http://www.rae.es

Schäfer, R. & F. Bildhauer (2013). *Web corpus construction*. (Synthesis Lectures on Human Language Technologies) San Francisco: Morgan & Claypool.

Suchomel, V. & J. Pomikálek (2012). Efficient web crawling for large text corpora. In A. Kilgarriff & S. Sharoff (Eds.), *Proceedings of the seventh web as corpus workshop (WAC7)* (pp. 39-43). Lyon, 2012.

Seretan, V. (2011). *Syntax-based collocation extraction*. (Text, speech and language technology series 44). Dordrecht: Springer.

Sinclair, J. (1966). Beginning the Study of Lexis. In C. Bazell, J. Catford, M. Halliday & R. Robins (Eds.), *In memory of J.R. Firth* (pp. 410-430). Longman: London.

Torres Cacoullos, R. & J. A. Walker (2011). Collocations in grammaticalization

and variation. In B. Heine & H. Narrog (Eds.), *Handbook of grammaticalization* (pp. 225-238). Oxford: Oxford University Press.

Tutin, A. (2008). For an extended definition of lexical collocations. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*. 1453-1460.

Williams, G. (2002). In search of representativity in specialised corpora – categorisation through collocation. *International Journal of Corpus Linguistics, 7* (1), 43-64.

Abstract

Language varieties should be taken into account in order to enhance fluency and naturalness of translated texts. In this paper we will examine the collocational verbal range for prima-facie translation equivalents of words like *decision* and *dilemma*, which in both languages denote the act or process of reaching a resolution after consideration, resolving a question or deciding something. We will be mainly concerned with diatopic variation in Spanish. To this end, we set out to develop a giga-token corpus-based protocol which includes a detailed and reproducible methodology sufficient to detect collocational peculiarities of transnational languages. To our knowledge, this is one of the first observational studies of this kind. The paper is organised as follows. Section 1 introduces some basic issues about the translation of collocations against the background of languages' anisomorphism. Section 2 provides a feature characterisation of collocations. Section 3 deals with the choice of corpora, corpus tools, nodes and patterns. Section 4 covers the automatic retrieval of the selected verb + noun (object) collocations in general Spanish and the co-existing national varieties. Special attention is paid to comparative results in terms of similarities and mismatches. Section 5 presents conclusions and outlines avenues of further research.

Keywords: collocation, diatopic varieties, translation, giga-token corpora

Author's address:

Gloria Corpas Pastor
Departmento de Traducción e Interpretación
Facultad de Filosofía y Letras
Campus de Teatinos s/n
Universidad de Málaga
29071-Malaga
Spain
gcorpas@uma.es